Weekly Report (16/11/2023) - Sindhuja Chaduvula

Unsupervised Learning Models and Dimensionality Reduction Techniques Applied on Vehicle dataset

Executive Summary:

The report presents the outcomes of applying advanced unsupervised learning techniques and dimensionality reduction methods to a roundabout vehicle dataset. The primary goal is to uncover inherent patterns and behaviors in the dataset, which consists of 900,000 Floating Car Data(FCD) across different attributes from 2000 vehicles.

Data Overview and Preprocessing:

The dataset captures a wide range of vehicle dynamics, including speed, acceleration, position, lane information for over 1500 seconds at a 0.1-second timestep. To enhance the analysis, the dataset underwent a preprocessing phase where data was aggregated by vehicle ID and type. This aggregation helped in distilling key features for each vehicle, incorporating statistical measures such as mean, variance and standard deviation.

Methodology:

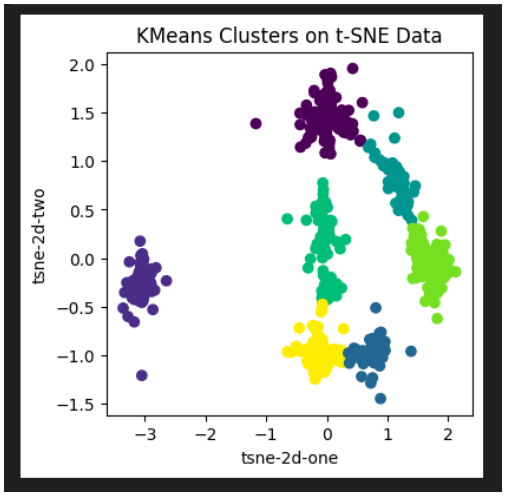
Our approach involved two primary stages:

1. **Dimensionality Reduction:**
    - **PCA**: Applied to project the data into lower-dimensional space while retaining most of the variance.
    - **t-SNE:** Used for its strength in visualizing high-dimensional data and preserving local structures.
    - **ICA**: A computational method for separating a multivariate signal into additive, independent non-Gaussian signals.
2. **Clustering Analysis:**
    - **KMeans:** This algorithm segmented the data into distinct clusters based on the reduced features.
    - **Hierarchical Clustering:** Offered a different perspective, revealing the nested structure of data groupings.

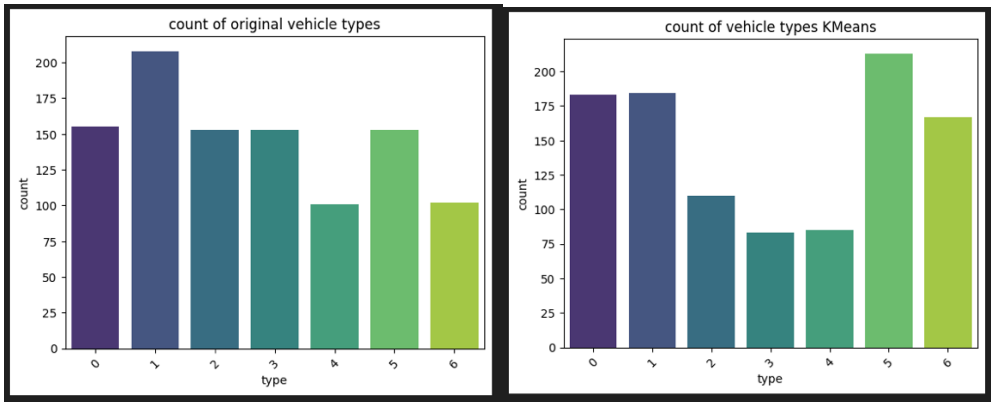
Findings and Analysis

T-SNE Data
KMeans

The comparison of vehicle type distributions between the original dataset and the clustering results obtained from KMeans applied to t-SNE transformed data reveals some interesting insights:



From the Analysis we can say that there's a shift in dominant categories, consistency and inconsistency in distribution, potential overlapping characteristics, implication of t-SNE transformation.

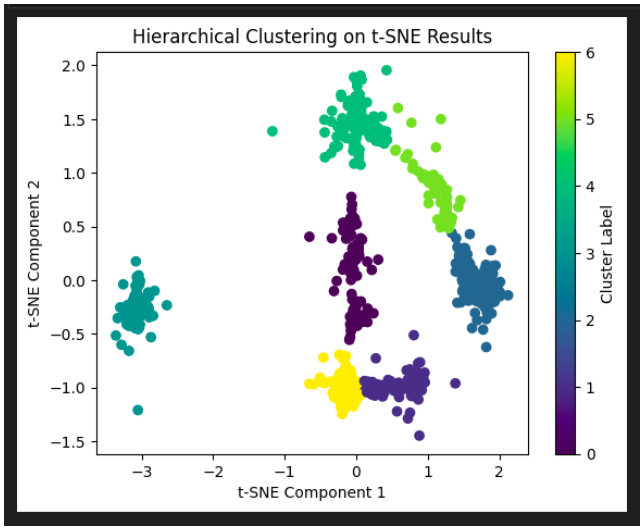


The comparison underscores the complexity and challenges in unsupervised learning, particularly in clustering high-dimensional data. The results from t-SNE and KMeans offer a different perspective on the dataset, which could be valuable for specific applications but might also require cautious interpretation to align with the expected distribution of vehicle types. This analysis suggests a need for further exploration, potentially considering other dimensionality reduction techniques, clustering algorithms, or a deeper dive into the feature engineering process to better understand and capture the inherent patterns in the vehicle data.

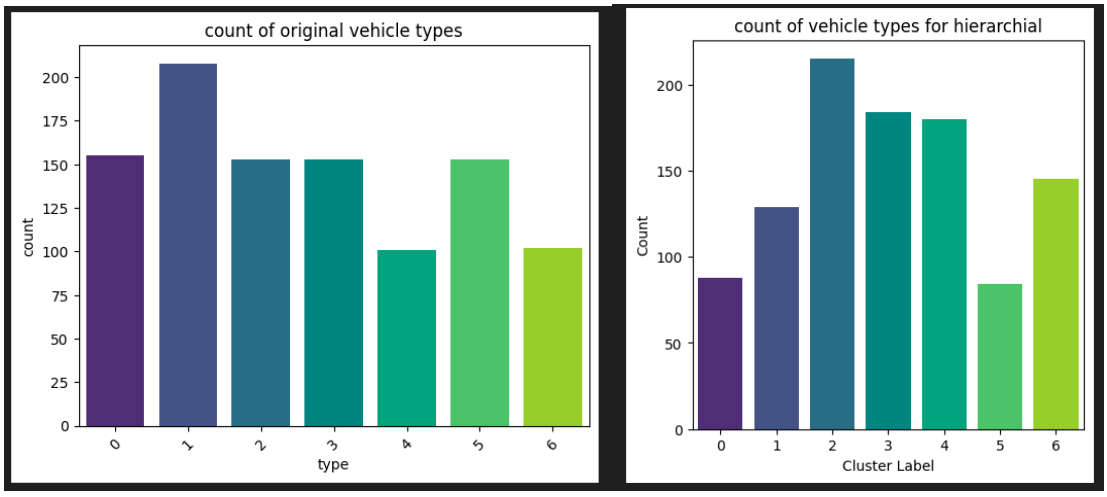
Hierarchical Clustering

To compare the distributions of vehicle types between the original dataset and the one obtained from t-SNE hierarchical clustering, we can focus on the changes in distribution patterns and what they may imply:

In the original dataset, Type 1 is the most common, type 4 and 6 are the least common. After applying t-SNE and hierarchical clustering the distribution shifts notably. Type 2 becomes the most common



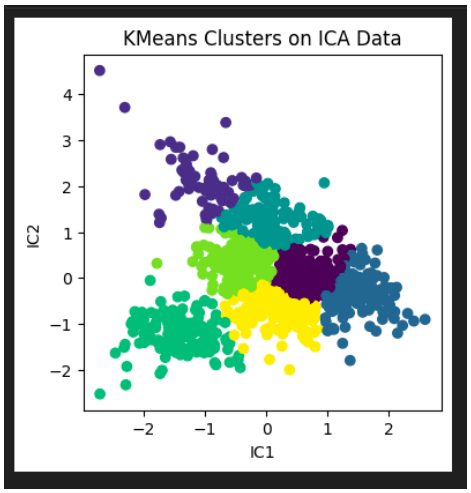
This is because of nature of t-SNE transformation, characteristics of hierarchical clustering, complexity and diversity in data. The hierarchical clustering results on the t-SNE transformed data reveal well-defined and distinct clusters.



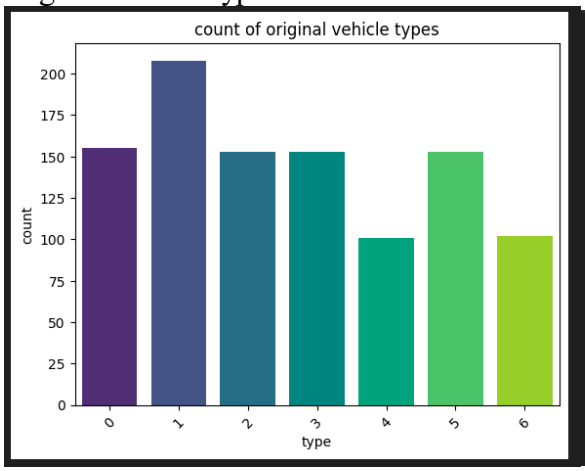
However, these clusters do not align perfectly with the original vehicle type distribution. This outcome highlights the influence of dimensionality reduction and clustering techniques on data representation and interpretation. It suggests that while these methods are effective in finding structures within the data, the resulting groupings may capture different aspects of vehicle behavior than the original categorizations. This analysis underscores the importance of understanding both the methodologies used and the nature of the data when interpreting unsupervised learning results. It also points to the potential for these techniques to uncover new patterns or relationships in the data that may not be apparent from the original vehicle types.

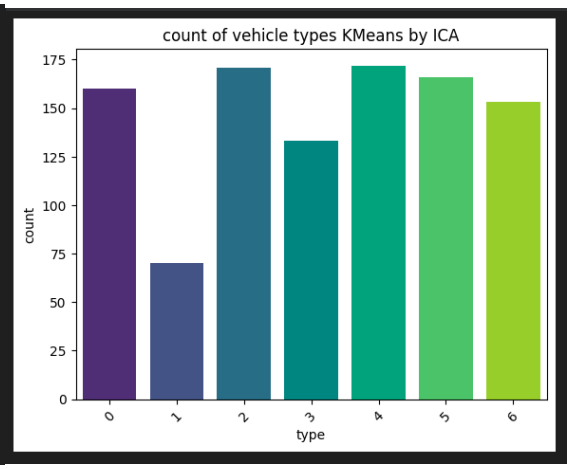
ICA
KMeans

To compare the distributions of vehicle types between the original dataset and the one obtained from t-SNE hierarchical clustering, we can focus on the changes in distribution patterns and what they may imply:



The changes in vehicle distribution resulting from ICA KMeans clustering, as compared to the original vehicle data, can be primarily attributed to the distinct methodological approach inherent in ICA. ICA aims to uncover latent factors in the data by isolating independent components that are statistically independent of one another. These components often represent hidden variables that are not directly observable in the raw dataset. For instance, in vehicular analysis, these could encapsulate nuanced attributes like driving styles or environmental conditions. Such components are combinations of original features and are adept at revealing underlying data structures that do not necessarily coincide with predefined vehicle type categories. Hence, when ICA is applied, it exposes new dimensions of the data, potentially correlating with different aspects of vehicle behavior than those defined by the original vehicle types.



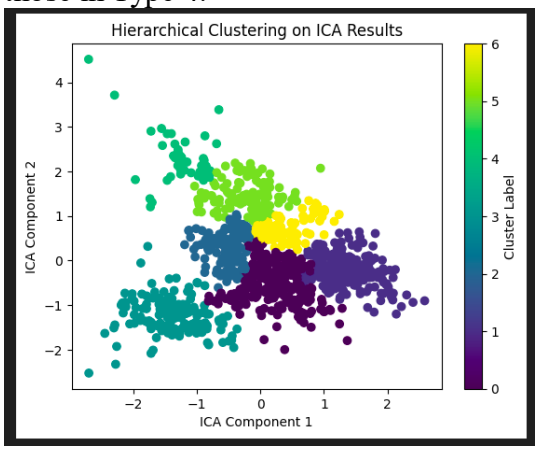


Post-ICA transformation, KMeans clustering groups vehicles based on these newly identified independent components, aligning vehicles with similar attributes into common clusters. This process can significantly alter the vehicle distribution across clusters, diverging from the original classification. For example, the prominent emergence of Type 4 vehicles in clusters post-ICA suggests that the behaviors characteristic of this type are more distinguishable through the lens of ICA. On the other hand, the dispersion of Type 1 vehicles across various clusters implies that the ICA components reflect a broader range of behaviors that dilute the

distinctiveness of this type. This redistribution underscores the complex, non-linear nature of the dataset and illustrates that clustering post-ICA reflects intrinsic statistical properties of the data, which may not always align with the original vehicle type labels. The clusters derived from this unsupervised learning approach offer valuable insights into new patterns within the data, highlighting the importance of contextual interpretation of these results, considering the transformative effect of the methods employed.

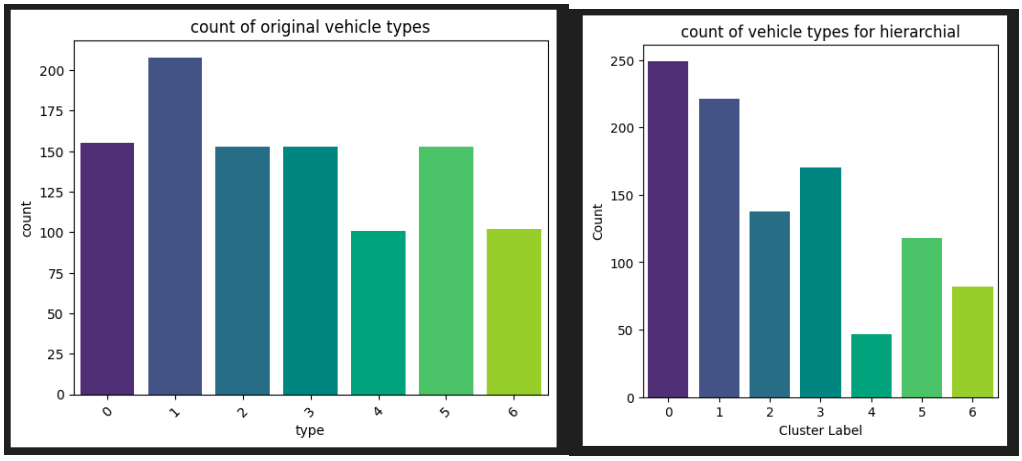
Hierarchical Clustering

The disparity between the original distribution of vehicle types and the distribution resulting from ICA hierarchical clustering is indicative of the complex transformations that the ICA method applies to the dataset. Independent Component Analysis aims to uncover hidden factors within the data by separating the dataset into components based on statistical independence, without consideration for the original vehicle type labels. These components may represent latent variables or inherent behavioral patterns that are not immediately apparent in the raw data. The ICA process can, therefore, highlight different aspects of the vehicles' operational characteristics, such as unique driving styles or responses to varying road conditions. In the vehicle distribution that emerges from ICA, we see a reordering of prevalence: Type 0 and Type 1 vehicles become more prominent, while Type 4 vehicles become the least common. This suggests that the behaviors or attributes most distinct in Types 0 and 1, as isolated by ICA, are more prevalent or consistent across the dataset than those in Type 4.



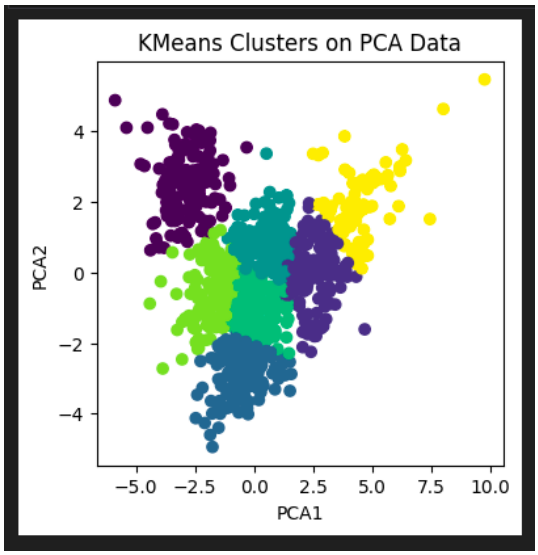
When hierarchical clustering is subsequently applied to the ICA-transformed data, the vehicles are grouped based on the new independent components rather than their original categorization. Hierarchical clustering constructs a tree-like model of vehicle groupings, revealing the relationships between data points as determined by their similarity in the ICA feature space. The resulting cluster distribution shows a significant shift: the largest cluster now corresponds to what was originally labeled as Type 0 vehicles, implying that, post-transformation, these vehicles share many common features that result in them being grouped together. Conversely, the distribution highlights a considerable reduction in the clustering of Type 4 vehicles, potentially indicating that the independent behaviors or characteristics for this type are more diffuse, leading to their broader spread across other clusters. These outcomes underscore the nuanced view that ICA hierarchical clustering provides, which may diverge from the original vehicle type assignments but offers a new perspective on the dataset, emphasizing statistical patterns and relationships potentially overlooked in the raw

data.



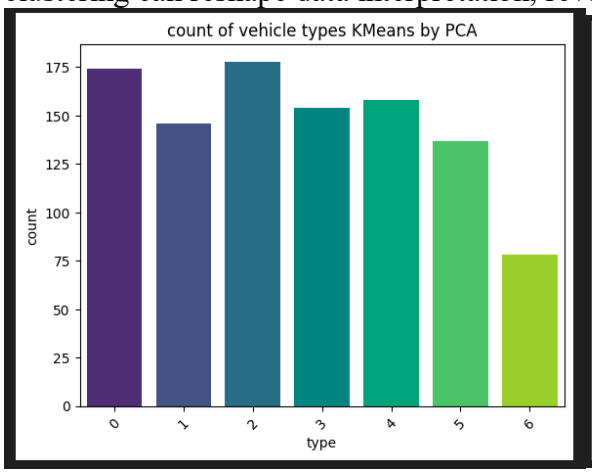
PCA
KMeans

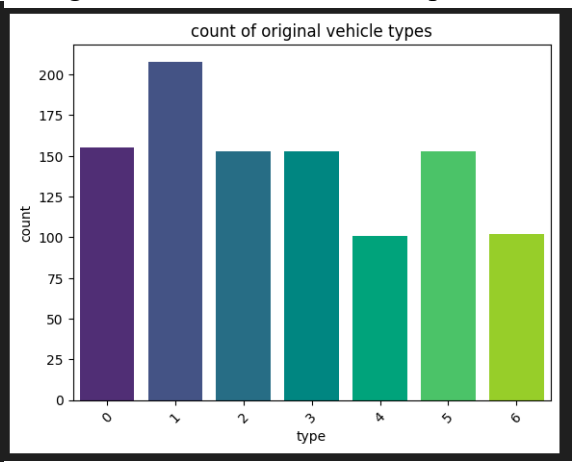
The shift in vehicle type distribution after PCA KMeans clustering reflects PCA's variance-focused dimensionality reduction, which may obscure finer original category distinctions. This linear technique projects data into principal components that most significantly explain its variance. When KMeans is applied to PCA-reduced data, it clusters vehicles based on these broad variance patterns. This can result in clusters that differ from original vehicle type frequencies, as PCA may merge subtle differences into larger variance trends, leading to new groupings based on the PCA features rather than the initial classifications.



In the original dataset, Type 1 vehicles were predominant, but after PCA and KMeans clustering, Type 2 becomes more prominent, suggesting PCA features align more with Type 2 characteristics. Types 0 and 4 also increase in cluster representation, hinting at their closer alignment with the principal variance components highlighted by PCA. Conversely, Types 1, 5, and 6 see reduced cluster representation, indicating that PCA may diminish their distinct features by blending them into general variance trends. This shift underscores how PCA and

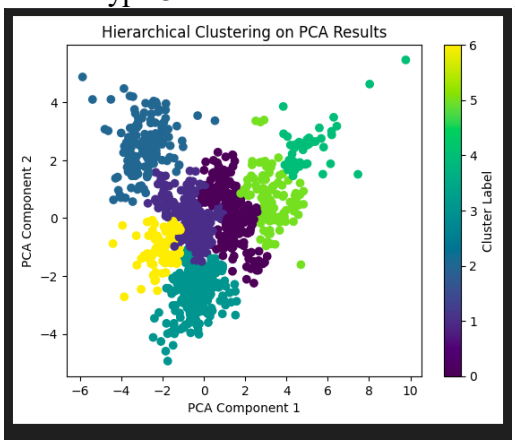
clustering can reshape data interpretation, revealing different vehicle behavior patterns.



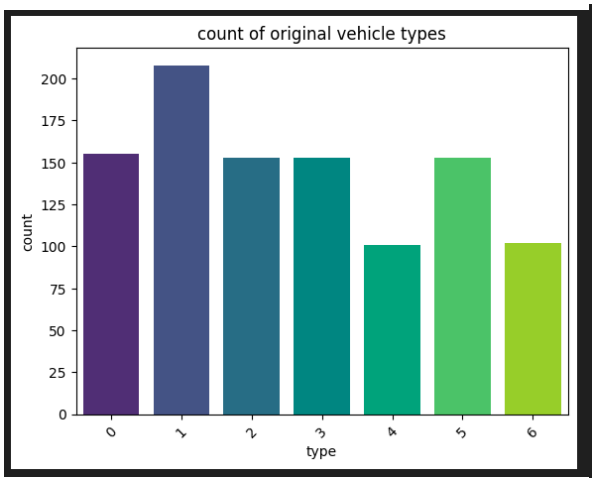


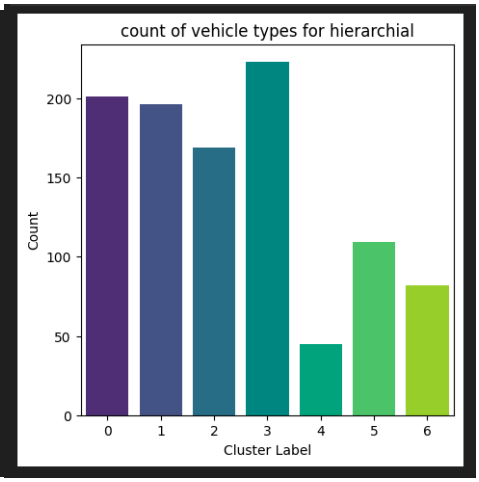
Hierarchical Clustering

The altered vehicle distribution post-PCA KMeans clustering, as compared to the original data, highlights the transformational effect of PCA on the dataset. PCA reduces dimensionality by projecting the data onto components that maximize variance, which doesn't necessarily align with the original vehicle type classifications. After this transformation, KMeans clustering forms groups based on the new feature space—where proximity is defined by the principal components rather than the original attributes. The redistribution, where Type 3 becomes the most prevalent post-clustering (223 instances) as opposed to Type 1 in the original data (208 instances), suggests that the PCA-derived features capture a structure in which Type 3's characteristics are more pronounced or common.



This shift in the clustering distribution also suggests that the original, nuanced features of certain vehicle types may have been overshadowed by broader patterns within the PCA-transformed space. For example, the original distinctiveness of Types 1, 5, and 6 seems diluted, with their cluster sizes reducing in the PCA-applied data. This can occur when the defining characteristics of these types are less aligned with the primary axes of variation highlighted by PCA. On the other hand, the decrease in representation for Type 4 vehicles (to 45 instances) in the clusters implies that PCA may have integrated their unique features into more general variance trends, causing them to be absorbed into larger clusters. The results underscore the influence of PCA on clustering outcomes, which can lead to a different understanding of data groupings and highlight the importance of considering the chosen dimensionality reduction technique's impact on subsequent analytical interpretations.


count of original vehicle types
count
type
0
1
2
3
4
5
6
0
25
50
75
100
125
150
175
200


count of vehicle types for hierarchial
Count
Cluster Label
0
1
2
3
4
5
6
0
50
100
150
200